早稲田大学大学院 理工学研究科

博 士 論 文 概 要

論 文 題 目

Stochastic Dynamical Systems with Hydropathy Index/Charge Trajectory for Transmembrane Secondary Structure Prediction

疎水性指標と電荷を軌跡とする確率力学系による膜タンパク質二次構造予測アルゴリズムの研究

申 請 者

鏑木	崇史
Takashi	Kaburagi

電気・情報生命専攻
学習型信号・情報処理システム研究

2008 年 10 月

20 世紀後半より始まったヒトゲノム解析を含め、生命現象を遺伝子・分子レベルで解き明かす研究群が行われてきた。そしてこれらの研究群は、タンパク質を構成するアミノ酸配列を数多く発見してきた。また、全国の総患者数が 14 万人を超えるパーキンソン病やアルツハイマー病をはじめとする神経変性疾患などに対する根源的な病理解析・病因解明に繋がるものとしても注目を浴びてきた。実際に現在に至るまで、ヒトゲノムの塩基配列の解読を達成し、ハンチントン舞踏病などの病理解明に貢献した。さらに、これらの研究群は、解読が進んだゲノムから生成される可能性のあるタンパク質のアミノ酸配列を数多く発見し、タンパク質機能と多くの神経変性疾患との関係を示唆するに至った。しかし、タンパク質は一般的にアミノ酸の配列だけでは機能せず、折りたたまれて特定の構造を構成してはじめて機能を持つ。このような中、分子レベルでの病理解析・病因解明をさらに進めるには、網羅的なタンパク質の構造解析が必須と考えられてきており、分子生物学・構造生物学の基本的なアプローチとなっている。近年、実験的にタンパク質の結晶構造が判明したものの数は飛躍的に伸びているが、アミノ酸配列が得られているタンパク質の数から比べるとまだ少ない。

一般的にタンパク質の構造の解明には大きく分けて３つのアプローチがあるといわれている。

（１）X 線結晶解析などを用いた実験的アプローチ：X 線結晶解析的アプローチでは目標となるタンパク質を精製、結晶化したうえ解析する。しかしタンパク質は結晶化に技術的、時間的、費用的に多くの困難が伴う事が多い。

（２）分子動力学計算を用いて解析するアプローチ：計算機を用いて目的のタンパク質の分子をモデル化し、その挙動を全原子レベルでシミュレーションするアプローチである。近年、計算機性能は飛躍的に向上しているが、比較的アミノ酸の数が小さいタンパク質であっても 2008 年現在の計算機能力では困難である事が多い。

（３）機械学習によるアプローチ：構造既知のデータから特徴を抽出し、構造未知のデータに対し予測を行うアプローチである。機械学習を用いるアプローチでは結晶化を必要とせず、また、分子動力学計算を実行するのに比べはるかに少ない計算量で予測を行う事ができる。機械学習アプローチが近年注目を浴びていきている背景のひとつがここにある。

本研究はタンパク質のうち特に膜タンパク質に着目し、機械学習的アプローチより膜タンパク質の構造を予測するアルゴリズムを構築することを目標とする。機械学習的手法は観測可能なデータを、背後に仮定されたモデルから発生したものと考え、そのモデルに付随するパラメタを推定して有用な規則、ルール、知識表現、判断基準などを抽出する手法である。観測されるデータには不確実性が含まれると考える。本論文を含め、不確実性は確率・統計の枠組みでとらえる事が多い。

膜タンパク質は細胞外の情報を細胞内に伝達するレセプタやイオンを透過させるイオンチャネルなど生命活動に極めて重要な役割を果たしている。

これらレセプタは、細胞に情報を伝達する物質（リガンド）と特異的に結合することにより、特定のシグナルを細胞内に伝達する。どのレセプタとどのリガンドがどのようにして細胞内に情報を伝達するかを解明することは、アルツハイマー病やパーキンソン病などの神経変性疾患などの病理解明、薬物応答の解明、さらにはゲノム創薬やオーダーメイド医療などの治療法確立に寄与するものと期待されている。その前段階とし、レセプタの構造を解明することは、対応するリガンドの特定、レセプタとの結合部位の解明、その動的挙動の解明に繋がると期待されている。

タンパク質の構造は一般的に一次から四次構造の４つの階層に分類がされている。一次構造とはアミ

ノ酸の 1 次元配列のことを指し、タンパク質をコードする遺伝子の塩基配列により直接決められている。アミノ酸配列が構成する比較的近距離の構造を二次構造と呼ぶ。これら二次構造がまとまって、折りたたまれることで三次構造を決めている。さらに、多くのタンパク質は三次構造を取ったアミノ酸配列が複数会合して形成されており、これらの立体構造を四次構造という。本研究で着目する膜タンパク質では生体膜に存在するという特別な環境が制約となり、三次構造が得られなくとも、二次構造が分かるだけでも機能の解明に有用であるとされている。本研究では特に二次構造に注目して、膜タンパク質の予測アルゴリズムの構築を目指す。

本研究で提案するアルゴリズムは、構造未知の膜タンパク質のアミノ酸配列を観測データとし、その膜貫通回数と膜貫通領域の推定を行うものである。膜タンパク質に限らず、一般にタンパク質立体構造予測問題には 3 つの重要な点が存在する。

(i)データは一次元の変数(C 末端からのアミノ酸数）に関して鎖状に連なっており、ある特定のアミノ酸は他のアミノ酸と空間的に関与している。

(ii)異なったアミノ酸配列が同様の構造を構成することがある。

(iii)学習に用いる事のできるデータセットは実験的に決定されている必要があるため、数が非常に限られている。

本研究では上記 3 つの問題を考慮し、確率・統計的な枠組みのモデルとそのパラメタ学習方法を提案する。

一般に、機械学習アルゴリズムの予測精度は 2 つの要素に依存すると考えられる:

(A)与えられた問題の特徴を的確にモデル構造に組み込むことが出来るか

(B)利用可能なデータセットから設計したモデルに付随するパラメタをいかに推定するか

ここで提案するアルゴリズムではモデルとして隠れマルコフモデル（Hidden Markov Model(HMM)）を用いている。HMM は逐次性を含むデータをモデル化するきわめて一般的な確率的ダイナミカルシステムの枠組みである。HMM が成功裏に適用されてきた問題の多くでは、時間に関する確率的ダイナミカルシステムであるのに対し、この研究を含めいくつかの生物データでは空間に関する確率的ダイナミカルシステムである。HMM は、内部に観測されない「状態」が存在すると仮定し、その状態に依存して「出力」が確率的に観測されると考えるモデルである。このような出力を確率的に扱う枠組みが上記の問題(ii)に対応する。内部の状態は確率的に別の状態に遷移する。このような隠れた状態列を仮定する枠組みが上記問題(i)に対応する。本研究で提案する HMM の出力は単なるアミノ酸配列の記号ではなく、疎水性指標と電荷等アミノ酸が固有に持つ物理化学量である。このような出力量の選択は問題(iii)解決に対する重要な要素のひとつとなっている　HMM を現実の問題に対して成功裏に適用するには適切な出力の選択、適切なトポロジー(状態の数や状態の間のグラフ構造)、および付随するパラメタを適切に学習するアルゴリズムが必要である。

本研究の特色は以下の通りである。

(1) 提案する HMM アルゴリズムでは、膜タンパク質を構成するアミノ酸配列が 20 種類の記号の連結であるとみなすのではなく、HMM の出力を疎水性指標と電荷から成る二次元二つの物理化学量の連なりであると考える。この定式化により 2 つのアミノ酸間の”距離”が定義でき、学習フェーズにおける過学習を防ぐ方法のひとつとしての”平滑化操作”を可能としている。平滑化操作は本課題のように学習データセットの数が少ない場合には重要である事が多い。先行研究では HMM を用いるアプ

ローチは行われているものの、これらのアプローチでは出力として物理化学量ではなく、アミノ酸を20 種類の記号として扱っているため距離を考慮する事ができず、平滑化が困難であった。
(2) 提案アルゴリズムではトポロジーとして開ループ構造（Left-to-Right）を仮定している。この構造は自己ループを除いては、閉ループが存在しないよう状態間の遷移が制約されている。より具体的には、状態が同じ状態に留まるか、次（右側）の状態に遷移するかの 2 通りに制約されている。このような特色のため、モデルに付随する状態遷移に関するパラメタの数が少なく、学習が容易になることが予想される。また、生物学的な視点から、このモデル構造を解釈することができる。開ループ構造では状態の遷移が初期状態から最終状態へと一方向にのみ遷移する。これはタンパク質を構成するアミノ酸の鎖がＮ末端からＣ末端へ一方向に連なっているというタンパク質がもつ特性をモデル化していると考えることが可能である。一方、先行研究での多くでは同じ状態に複数回遷移可能となる閉ループ構造が仮定されている。
(3) 本研究では学習に用いることのできる膜タンパク質のデータセットの少なさを考慮した学習アルゴリズムを提案している。HMM パラメタの学習にはバウム・ウェルチアルゴリズム(Baum-Welch Algorithm)が用いられる事が多いが、提案手法では用いていない。バウム・ウェルチアルゴリズムは局所最適解に陥る可能性があるためである。特に学習データセットの数が十分ではない場合には最適解到達が困難である事が多いと予想される。
本論文の構成は以下の通りである。
第 1 章で、本研究の背景について述べる。まず本研究の対象としている膜タンパク質の概要について述べる。次に膜タンパク質の構造解析の目的を述べた後、どのようなアプローチが行われているかについて、実験的な手法、計算機的な手法の両面から述べる。特に本研究と密接に関連するバイオインフォマティクスによるアプローチについては詳しく述べる。

第 2 章では提案する基本モデル、ならびに基本アルゴリズムについて述べる。前述の通り、本研究で提案する HMM では疎水性指標と電荷の 2 次元物理化学量を出力としている。この章では 2 次元出力を持つ HMM の定式化、Left-to-right 構造をもつ HMM トポロジーについて、さらに HMM に付随するパラメタの学習方法について述べる。学習された HMM を用いて構造未知膜タンパク質のアミノ酸配列が与えられたときの膜タンパク質構造推定の方法について述べる。より具体的には膜貫通回数予測の方法、さらに膜貫通領域の推定アルゴリズムについて述べる。提案する基本モデルは公開されている膜タンパク質データベースを用いて評価を行う。また、既存のアルゴリズムについても同じデータセットを用いてその評価を試み、報告する。
第 3 章では第 2 章にて述べた基本モデルと基本アルゴリズムを拡張し、2 つの方法で予測精度の向上を目指している。具体的には、領域推定ステップを改良する方法とモデル構造を改良する方法の 2 つを議論する。本研究で提案しているアルゴリズムの領域推定ステップは HMM の状態推定にて行われている。この状態推定方法はいくつか考えることができ、それらの比較検討を行っている。また、モデル構造の改良では従来の HMM を一般化した一般化隠れマルコフモデル（Generalized Hidden Markov Model（GHMM））を用いて膜タンパク質をモデル化することを試みている。本論文では GHMM と従来の HMM との比較を行う。

第 4 章では上記第 2 章と 3 章をまとめ、本研究の今後の展望について述べる。